WSS TALK 10/27/93

USING DIFFERENT PRECIPITATION TERMS TO FORECAST CORN AND SOYBEAN YIELDS

M. Denice McCormick
USDA/NASS/Research Division/3251 Old Lee Hwy., Room 305, Fairfax, VA. 22030

KEY WORDS: Precipitation, regression models

INTRODUCTION

In 1990, the National Agricultural Statistics Service (NASS) introduced new models to forecast yield for corn and soybeans on the regional and State levels in a plan to phase out the older, less accurate models (Birkett 1990). An annual survey collects data from randomly selected sample plots in randomly selected fields. The old regression models predicted the components of yield such as number of pods per plant and weight per pod at the plot level based on five years of previous data. Plot level data were then aggregated to the State level. The new models are also regression models, and have initially been developed to predict yield directly rather than the components of yield using survey data aggregated to the regional level. Regions are constructed from the set all States that participate in the annual survey. A longer period of years in the historic data set must be used since only one data point is used to represent each year.

McCormick and Birkett (1992) tried to improve the accuracy of early season soybean yield forecasts by adding a term that represented total accumulated precipitation throughout the growing season from April 1 until the forecast date at a six-State regional level. The analysis indicated that soybean forecast accuracy at the regional level was not improved using this particular term. Based upon this result, two recommendations were made. One was to evaluate alternative time frame terms, such as monthly precipitation totals. The other was to use them to forecast other major agricultural crop yields. This paper reports results when separate monthly precipitation terms were added to corn and soybean yield forecast models. It considers data for thirteen years, 1980 to 1992. The soybean States included in the study are Arkansas, Illinois, Indiana, Iowa, Missouri, Minnesota, Nebraska, and Ohio. The corn States are Illinois, Indiana, Iowa, Michigan, Minnesota, Missouri, Nebraska, Ohio, South Dakota, and Wisconsin. The performance of each model is compared to official operational model performance.

This study evaluates multiple regression models which use precipitation and survey variables to forecast end-of-season crop yields. In previous research, the models showed improved performance using aggregated survey variables at the regional level. Therefore, this method was also used to aggregate the precipitation variables.

DATA

Precipitation Data

Precipitation variables used in the models represent total precipitation for a particular month at the regional level. The data are provided from a network of National Weather Service weather stations in each State. The variable is constructed as follows:

$$P_t = \frac{\sum_{s=1}^{S} A_{ts} R_{ts}}{\sum_{s=1}^{S} A_{ts}}, \qquad (1)$$

where

P_t = the average total precipitation within selected month for the region for year t,
S = the number of States covered,
A_{ts} = the acres for harvest for year t, State s, and
R_{ts} = the average total precipitation within selected month for year t, State s,

where

$$R_{ts} = \frac{\sum_{d=1}^{D_s} A_{tsd} E_{tsd}}{\sum_{d=1}^{D_s} A_{tsd}},$$

A_{tsd} = the acres for harvest for year t, State s, district d, and
D_s = the number of districts per State s,
E_{tsd} = the average total precipitation within selected month for year t, State s, district d,

where

W_{tsd} = number of weather stations for year t,

$$E_{tsd} = \frac{1}{W_{tsd}} \sum_{w=1}^{W_{tsd}} U_{tsdw}$$

State s, district d, and

U_{tsdw} = total precipitation within selected month for year t, State s, district d, weather station w.

Survey Data

The construction of the independent variables for the regional regression models for both soybeans and corn is discussed by Birkett (1990, 1993). For soybeans for the month of August, the independent variable (Z_t) is the estimated number of lateral branches per eighteen square feet. For September, the independent variable is the estimated number pods with beans per eighteen square feet. These regional-level estimates for soybeans are constructed as follows:

$$Z_t = \frac{\sum_{s=1}^{S} A_{ts} F_{ts}}{\sum_{s=1}^{S} A_{ts}} \qquad (2)$$

where

A_{ts} = the acres for harvest for year t, State s, and

F_{ts} = number of lateral branches per 18 sq. feet year t, State s,

$$F_{ts} = \frac{1}{m_{ts}} \sum_{j=1}^{m_{ts}} B_{tsj} L_{tsj} ,$$

where

m_{ts} = the number of samples in J_{ts} year t, State s,

J_{ts} = the subset of samples classified in maturity categories 2-6 (or 1-6 in southern States), year t, State s,

B_{tsj} = plants per 18 square feet for year t, State s, sample j,

L_{tsj} = lateral branches per plant year t, State s, sample j (for August) or

= estimated pods with beans per plant per 18 sq. feet, year t, State s, sample j (for September).

Corn independent variables (Z_t) are more complex as they are a function of both plant counts and average kernel row length per square foot. C_{ts} is substituted for F_{ts} in equation (2). In August, it is calculated as:

$$C_{ts} = \frac{1}{m_{ts}} \sum_{j=1}^{m_{ts}} (U_{tsj} + V_{tsj}) \bar{K}_{tsj} ,$$

where

C_{ts} = a function of the number of stalks with ears, the number of ears with kernels, and the average kernel row length per square foot,

U_{tsj} = number of stalks with ears per sq. ft., year t, State s, sample j,

V_{tsj} = number of ears with kernels per sq. ft., year t, State s, sample j, and

$\bar{K}_{tsj}$ = the average kernel row length per ear, year t, State s, sample j.

In September, C_{ts} is calculated as:

$$C_{ts} = \frac{1}{m_{ts}} \sum_{j=1}^{m_{ts}} (V_{tsj}) \bar{K}_{tsj} .$$

For both forecasts, data are used from the subset of samples in maturity categories 3-6 for year t, State s.

Yield Data

The regional yield values included in this study were calculated as follows:

$$Y_t = \frac{\sum_{s=1}^{S} A_{ts} Y_{ts}}{\sum_{s=1}^{S} A_{ts}} , \qquad (3)$$

where

Y_t = final regional yield for year t, and

Y_{ts} = NASS State yield year t, State s.

METHODOLOGY

Regression analysis was used to evaluate the performance of precipitation data in combination with

survey data. Multiple linear regression models with associated diagnostics for model fit and forecast accuracy were examined. The basic regression models analyzed were:

1: $Y_t = \beta_o + \beta_1 Z_t + \epsilon_t$

2: $Y_t = \beta_o + \beta_1 Z_t + \beta_2 Z_t^2 + \epsilon_t$

3: $Y_t = \beta_o + \beta_1 Z_t + \beta_2 P_t + \epsilon_t$

4: $Y_t = \beta_o + \beta_1 Z_t + \beta_2 Z_t^2 + \beta_3 P_t + \epsilon_t$.

Model 2 is the official model used by NASS to forecast August corn and soybeans and September soybeans. However, Model 1 is the official model used to forecast September corn. Models 3 and 4 use one monthly precipitation term. Analysis was conducted to determine which month from the growing season provided optimal forecasting capability. Also, models with multiple monthly precipitation terms were examined.

Model Evaluation Criteria

The primary model evaluation criterium is the set of prediction intervals (PI) for the minimum, median, and maximum yielding years over 13 years in the study. For soybeans, these years were 1988, 1981 and 1990, and for corn, they were 1983, 1989 and 1992, respectively. A second criterium is the adjusted coefficient of determination, R_a^2 which provides a measure of correspondence between predicted and actual yields. Both the PI and R_a^2 are based on the sum of squared differences from the least squares analysis used to derive the model parameters.

1. The prediction interval (PI) refers to half the confidence interval length for the predicted value of a future Y for a given future year o. That is, at the α significance level,

$$PI = t(1-\frac{\alpha}{2};n-1-p)SD(\hat{Y}_o),$$

where

$$SD(\hat{Y}_o)=s[(x_o'(X_o'X_o)^{-1}x_o) + 1]^{\frac{1}{2}},$$

s = (residual MSE)$^{1/2}$,
x_o = relevant p-dimensional row vector of independent variables for year o (for example, in Model 3: p= 3, x_o = [1, Z_o, P_o]),
X_o = relevant (n-1 x p) matrix of independent variables (excludes x_o),
n = number of years, and
p = number of parameters.

The X_o matrix excludes the row vector x_o, so that the PI reflects the accuracy expected in an operational model where current year data are not included in the model development. A significance level of 0.32 was used for this study, which provides t values near 1.0. Consequently, the future Y will fall within the calculated PI of the predicted Y approximately 68% of the time.

2. R_a^2 is used as a goodness-of-fit test for each model with an adjustment made for the corresponding degrees of freedom (Draper and Smith 1981).

R_a^2 is calculated as:

$$R_a^2 = 1-\frac{(RSS_p)/(n-p)}{(CTSS)/(n-1)},$$

where

RSS_p = the residual sum of squares taking the changing number of parameters into account,
CTSS = the corrected total sum of squares,
n = the number of years, and
p = the number of parameters.

Outlier Identification

Since the purpose of the models is to make forecasts, the rstudent statistic (also called the studentized residual) was used to help identify outliers to be excluded from the model. This statistic was recommended in Belsley, Kuh and Welsh (1980). It is similar to the standardized residual:

$$r_{si} = \frac{r_i}{s\sqrt{1-h_i}},$$

where

r_i = i^{th} residual,
s = (residual MSE)$^{1/2}$, and
h_i = $x_i'(X'X)^{-1}x_i$.

Here, s is replaced by s(i). S(i) is the estimate of σ with the i^{th} observation deleted. In a forecasting model, rstudent measures how many prediction standard errors the forecast is from the observed Y. Observations with absolute values of rstudent greater than 3.0 were identified as outliers. The rstudent statistic is distributed closely to the t-distribution with n-p-1 degrees of freedom.

RESULTS

Regression analysis was conducted on a number of different models using different monthly precipitation terms. Tables 1 and 2 present the prediction intervals and R_a^2 for the official linear or quadratic model using survey data only and then results adding the optimal monthly precipitation term. In both tables, the prediction intervals relate to the years with minimum, median, and maximum regional yields.

Table 1: August Results

Model	R_a^2	Prediction Intervals		
		min	med	max
CORN:				
Official	.87	7.0	5.7	6.2
P_t=July	.93	5.4	4.3	4.9
SOYBEANS:				
Official	.70	2.8	2.3	2.7
P_t=July	.74	2.3	2.1	2.3

Note: August corn: both models have outlier year 1988 removed.

Table 2: September Results

Model	R_a^2	Prediction Intervals		
		min	med	max
CORN:				
Official	.97	3.6	3.2	3.4
P_t=June	.98	2.3	2.0	2.2
SOYBEANS:				
Official	.89	1.7	1.6	1.6
P_t=August	.88	1.9	1.7	1.9

Note: September corn: Official model removed 1990; Precip model removed 1988.

CONCLUSIONS

Except for the September soybean forecast, the precipitation models performed better than the official forecast models since their prediction intervals were consistently smaller. Contrary to previous indications, the August forecast models demonstrated that the addition of a monthly precipitation term with a survey term does improve forecasts for both crops. For both periods, the corn forecast seemed to benefit the greatest. There is no evidence that a change from the official model is warranted for September soybeans.

BIBLIOGRAPHY

Belsley, David A, Kuh, Edwin, Welsh, R.E., (1980), Regression Diagnostics, John Wiley & Sons.

Birkett, Thomas R., (1990) "The New Objective Yield Models for Corn and Soybeans", SMB Staff Report Number SMB-90-02, USDA.

Birkett, Thomas R., (1993) "Yield Models for Corn and Soybeans Based on Survey Data", USDA, Proceedings, ICES Conference, Buffalo, New York.

Draper, N.R., Smith, H., (1981), Applied Regression Analysis, John Wiley & Sons Second Edition.

McCormick, M. Denice, Birkett, Thomas R. (1992) "Evaluating the Addition of Weather Data to Survey Data to Forecast Soybean Yields", SRB Research Report No. SRB 92-11, USDA.